Feature Extraction Techniques
CMU at TRECVID 2004

Ming-yu Chen and Jun Yang

School of Computer Science
Carnegie Mellon University
Outline

• Low level features
• Generic high level feature extractions
  • Uni-modal
  • Multi-modal
  • Multi-concept
• Specialized approach for person finding
• Failure analysis
Low level features overview

• Low level features
  • CMU distributed 16 feature sets available to all TRECVID participants
  • Development set: http://lastchance.inf.cs.cmu.edu/trec04/devFeat/
  • Test set: http://lastchance.inf.cs.cmu.edu/trec04/testFeat/
  • These features were used for all our submissions
  • We encourage people to compare against these features to eliminate confusion about better features vs better algorithms
Low level features

• Image features
  • Color histogram
  • Texture
  • Edge

• Audio features
  • FFT
  • MFCC

• Motion features
  • Kinetic energy
  • Optical flow

• Detector features
  • Face detection
  • VOCR detection
Image features

- 5 by 5 grids for key-frame per shot
- Color histogram (*.hsv, *.hvc, *.rgb)
  - 5 by 5, 125 bins color histogram
  - HSV, HVC, and RGB color space
  - 3125 dimensions (5*5*125)
  - row-wise grids
  - 19980202_CNN.hsv

- Texture (*.texture_5x5_bhs)
  - Six orientated Gabor filters

- Edge (*.cannyedge)
  - Canny edge detector, 8 orientations
Audio features & Motion features

- Every 20 msecs (512 windows at 44100 HZ sampling rate)
  - FFT (*.FFT) – Short Time Fourier Transform
  - MFCC (*.MFCC) – Mel-Frequency cepstral coefficients
  - SFFT (*.SFFT) – simplified FFT

- Kinetic energy (*.kemotion)
  - Capture the pixel difference between frames

- Mpeg motion (*.mpgmotion)
  - Mpeg motion vector extracted from p-frame

- Optical flow (*.opmotion)
  - Capture optical flow in each grid
Detector features

- Face detector (*.faceinfo)
  - Detecting faces in the images

- VOOCR detector (*.vocrinfo and *.mpg.txt)
  - Detecting and recognizing VOOCR
Closed caption alignment and Shot mapping

- Closed caption alignment (*.wordtime)
  - Each word in the closed caption file is assigned an approximate time in millisecond
- Shot Break (*.shots)
  - Provides the mapping table of the shot

- We encourage people to utilize these features
  - Eliminate confusion of better features or better algorithms
  - Encourage more participants who can emphasize their efforts on algorithms
Outline

- Low level features
- Generic high level feature extractions
  - Uni-modal
  - Multi-modal
  - Multi-concept
- Specialized approach for person finding
- Failure analysis
Generic high level feature extraction

Uni-Modal Features
- Structural Info.
  - Timing
- Textual Info.
  - Transcript
- Audio Info.
  - SFFT
  - MFCC
- Visual Info.
  - Video OCR
  - Face Feature
  - Kinetic Motion
  - Optical Motion
  - Gabor Texture
  - Canny Edge
  - HSV/HVC/GRB Color

SVM-based Combination

Multi-modal Features
- Concept 1
- Concept 2
- Concept 3
- Concept 4

Multi-concepts Combination
- Concept 168

Feature Tasks
1. Boat / Ship
2. Madeleine Albright
3. Bill Clinton
10. Road

Common Annotation

Carnegie Mellon
Multi-concepts

- Learning Bayesian Networks from 168 common annotation concepts
- Pick top 4 most related concepts to combine with the target concept

<table>
<thead>
<tr>
<th>Multi-concept</th>
<th>Combined Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boat/Ship</td>
<td>Boat, Water_Body, Sky, Cloud</td>
</tr>
<tr>
<td>Train</td>
<td>Car_Crash, Man_Made_scene, Smoke, Road</td>
</tr>
<tr>
<td>Beach</td>
<td>Sky, Water_Body, Nature_Non-Vegetation, Cloud</td>
</tr>
<tr>
<td>Basket Scored</td>
<td>Crowd, People, Running, Non-Studio_Setting</td>
</tr>
<tr>
<td>Airplane Takeoff</td>
<td>Airplane, Sky, Smoke, Space_Vehicle_Launch</td>
</tr>
<tr>
<td>People Walking/running</td>
<td>Walking, Running, People, Person</td>
</tr>
<tr>
<td>Physical violence</td>
<td>Gun_Shot, Building, Gun, Explosion</td>
</tr>
<tr>
<td>Road</td>
<td>Car, Road_Traffic, Truck, Vehicle_Noise</td>
</tr>
</tbody>
</table>
### Top result for TRECVID tasks

- Uni-modal gets 2 best over CMU results
- Multi-modal gets 3 best, but includes Boat/Ship which is the best overall all
- Multi-concept gets 6 best

<table>
<thead>
<tr>
<th></th>
<th>Boat*</th>
<th>Train</th>
<th>Beach</th>
<th>Basket</th>
<th>Airplane</th>
<th>Walking</th>
<th>Violence</th>
<th>Road</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni-modal</td>
<td>0.097</td>
<td>0.001</td>
<td>0.013</td>
<td>0.503</td>
<td>0.021</td>
<td>0.015</td>
<td>0.005</td>
<td>0.036</td>
</tr>
<tr>
<td>Multi-modal</td>
<td>0.137</td>
<td>0.001</td>
<td>0.023</td>
<td>0.517</td>
<td>0.014</td>
<td>0.008</td>
<td>0.002</td>
<td>0.045</td>
</tr>
<tr>
<td>Multi-concept</td>
<td>0.110</td>
<td>0.001</td>
<td>0.039</td>
<td>0.517</td>
<td>0.035</td>
<td>0.099</td>
<td>0.003</td>
<td>0.062</td>
</tr>
</tbody>
</table>
Outline

- Low-level features
- Generic high-level feature extractions
- Specialized approaches for person finding
- Failure analysis
A Text Retrieval Approach

• Search for shots with names appearing in transcript
  • Vector-based IR model with TF*IDF weighting

• Temporal Mismatch Drawback
  • Faces do not always temporally co-occur with names
  • Cause false alarms and misses
Expand the Text Retrieval Results

- Propagate the text score to neighbor shots based on the distribution
- \( \text{Timing score} = F(\text{Distance (shot, name)}) \)

Model 1: Gaussian model (trained using Maximum Likelihood)
Model 2: Linear model (different gradients set on two sides)
Context Information

- Sometimes, a news story has the name but not the face
  - E.g., “.... a big pressure on Clinton administration ...”
  - Cause many false alarms
  - Related to the context “Clinton administration”

- Given the context, how likely a story has the face?
  - Collect bigrams of type “___ Clinton” or “ Clinton ___”
  - Compute $P_i (\text{Clinton appears in the story} \mid \text{bigram}_i)$

<table>
<thead>
<tr>
<th>Bigram</th>
<th>$P (\text{face} \mid \text{bigram})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton says</td>
<td>0.627474</td>
</tr>
<tr>
<td>Clinton made</td>
<td>0.625652</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
<tr>
<td>Clinton administration</td>
<td>0.242973</td>
</tr>
</tbody>
</table>
Multimodal Features

• Multimodal features provide weak clues for person search
  • Face detection – shots w/o detected faces are less likely to be the results
  • Facial recognition – matching detected faces with facial model based on Eigenface representation
  • Anchor classifier – anchor shots rarely have intended faces
  • Video OCR – Fuzzy match by edit distance between video OCR and the target name
Combining Multimodal Info. with Text Search

• Updated Text Score: \[ R' = R \times \text{Timing Score} \times \text{Avg} (P_{\text{bigram}_i}) \]

• Linear combination of all the features with text score
  • Features normalized into (pseudo-) probabilities \([0,1]\)
  • Feature selection based on chi-square statistics
  • Combinational weights trained by logistic regression

<table>
<thead>
<tr>
<th>Features</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Updated text score</td>
<td>6.14</td>
</tr>
<tr>
<td>Face similarity</td>
<td>3.94</td>
</tr>
<tr>
<td>Face detection</td>
<td>0.50</td>
</tr>
<tr>
<td>Anchor detection</td>
<td>-5.65</td>
</tr>
</tbody>
</table>
• One of our “Albright” runs is the best among all submissions
• Combining multimodal features helps both tasks
• Context helps “Clinton” but hurts “Albright”
  • Probably due to sparse training data for “Albright”
Outline

- Low-level features
- Generic high-level feature extractions
- Specialized approaches for person finding
- Failure analysis
Performance Drop on Anchor Classifier

- 95% cross-validation accuracy on development set
- 10 videos in testing set have 0 or 1 detected anchor
- Average # of anchor shots per video is 20-30
Different Data Distribution

- Different images: change on background, anchor, clothes
  - Common types (development set)
  - Outliers (testing set)

- Similar images, but probably different MPEG encoding
  - “Peter Jennings” has similar clothes and background in both sets
  - In videos with “Peter Jennings” as the anchor
    - 19 detected per video in development set
    - 13 detected per video in testing set
Other Semantic Features

• Similar performance drop observed on Commercial, Sport News, etc
  • Compromises both high-level feature extraction and search

• Possible solutions
  • Get consistent data next year
  • Rely less on sensitive image features (color, etc)
  • Rely more on robust features -- “Video grammar”
    - Timing, e.g., sports news are in the same temporal session
    - Story structure, e.g., the first shot of a story is usually an anchor shot
    - Speech, e.g., anchor’s speaker ID is easy to identify
    - Constraints, e.g., weather forecast appear only once
  • Re-encoding MPEG video
Thank you!